（招待発表）

Python と TwitterAPI によるビッグデータ事始め

荒川　歩(武蔵野美術大学)

1. はじめに

社会言語科学では，従来，少数の事例を丁寧に質的に分析する方法と，大量のデータを統計的に分析する方法というお互いに利点と欠点のある方法が相互に補い合うように用いられてきた．このうち，前者の質的な分析は，研究者の人力に依存よる部分が多く，分析の背後にある方法論の精緻化という形で発展をしているのに対して，後者の一部は，大量のデータを貯蔵し，効率的に検索できるコンピュータや，言語データを機械的に処理するテキストマイニングソフトウェアなど技術的な発展の影響を受けてきた．それでも従来の量的研究が対象とするデータは，コーパスに代表されるように，出処は明確であるが量的に限られるものであった．ところが近年，インターネットの普及により，Twitter などの（位置情報などさまざまな情報と紐付いてはいることもあるが）誰がどのような文脈で発したかの情報には欠ける極めて大量の言語データ，いわゆるビッグデータが分析の対象として用いられるようになった．

Twitter とは，140 字という制限字数の中で匿名で簡便に情報を不特定多数に発信したり，それを読んだり，共有したりすることのできるインターネット上のサービスであり，フォローとよばれる機能を相互に使うことで，他のユーザーに自身の現在の状態や関心を伝え，また，友人や有名人の関心や情報をリアルタイムで得ることができるものである．この Twitter のデータを言語データとして利用することについては，「言語資料としてどのような性格を有しているのか，実は未だ詳らかではない」という指摘(岡田・西川, 2016)があり，その使われ方は，時間と伴に変化し，もともとの独り言的に表出される私的発話から，社会的発話の割合が増えつつあるという指摘もある(澤山・三宅, 2018)．また，十分な利用者の情報に欠けているため利用者の特徴を含めた多面的な分析をする必要があるという指摘もある(北村・河井・佐々木, 2017)．しかし，ツイートはある種の言語使用であると考えられることから，これを使うことで，大きな標本を対象に，ある言葉が地域的にどの地名と結びついて使われているか，あるいは，ある言葉と同時に使われる言葉が時間的にどのように変化をするかを巨視的に調べることができる．

そのため，現在でもさまざまな観点から，言葉の変化を捉える手段として，Twitter が分析の対象として用いられている．例えば，五味・辰巳・新田(2012)は，「違う」の形容詞化について調査し，後に続く語としては「違わない・違(う)くない」と続く場合の形容詞化率が最も高いことを指摘した．また，全く別の種類の研究として，保田・岡本・荒牧(2012)は，ファッション雑誌等で，読者に新しい印象を与えるために，語を組み替えて使用する現象に着目し，雑誌における表現について分析するとともに，Twitter において人びとが使う言語表現や用法にも着目し，キーワード検索を用いて「大人」「ｘ＋」「大人＋ｘ」「大人な」といった用法について調査を行い，その結果として「『大人なガーリッシュ』や『大人女子』のような新たな用法が現われることで，『大人』用例の割合に変化が現われている」と指摘している．

しかし現状では，Twitter というビッグデータの分析にはプログラミング等の知識を要するため，それらの専門的な知識をもった人以外には障壁が大きいと思われる．そこで本研究では，素人として大量のツイートデータを収集し，テキストマイニングで簡単な分析を行う方法について紹介する．例として，近年「大人コーデ」や「双子コーデ」といったように利用が増えている「コーデ」と言う言葉に注目し，それがどのような言葉と同時にツイートされているかについて検討を行う．

2. ツイートの取得の方法

ツイートは，Twitter 社が管理しており，ツイートの取得方法は，さまざまな方法が提供されている．無料で利用できる公式の方法は，Twitter 社が提供する Standard Search API というウェブ上のシステムに，Perl や Python などで作成したプログラムでアクセスして情報を得るものである．Standard Search API は，緯度経度情報を含んだツイートの検索も可能であり，これは社会言語学的には有用な情報だと考えられるが，この情報が含まれているものはツイートのごく一

部である．また，検索可能な対象期間が公式には7日間であり，一度の検索で引き出すことのできる件数と一定時間内に検索を行うことのできる回数がかなり限られている．なお，このほか，より制限が緩やかな有料サービスも提供されている．

ここでは，わかりやすいことで，近年，利用が伸びているプログラム言語である Python を用い，Standard Search API にアクセスして，ツイートを収集することを試みる．

2.1 API キーの取得

API を利用するには利用者を識別するための Consumer Key, Consumer Secret, Access Token, Access Token Secret の 4 つが必要である．Twitter アカウントを取得したうえで，https://apps.twitter.com/ にアクセスして，指示に従って，上記の４つのキーを取得する．

2.2 Python のインストール

https://www.python.org/ にアクセスして，Python をインストールする．Windows よりも Linux や Mac のほうが，形態素分析に用いる MeCab との相性がいいという指摘もある．インストールをする際，どのディレクトリからでも Python が使えるように Path を設定するかにチェックを入れる欄があるのでチェックを入れる．

2.3 ツイートの収集

http://www.mivurix.sakura.ne.jp/mivurix/twitter.html から実際に，「コーデ」を含むツイートを収集する際に用いたプログラムをダウンロードできる．プログラムをメモ帳などで開き，検索ワードや CONSUMER_KEY などの部分を自分の API キーに置き換え，Windows のコマンドプロンプトでプログラムのディレクトリに移動して「pyhon プログラム名」と入力すれば，時間をかけて，test.html というファイルに収集したツイートが書き込まれる(すべてのファイルは utf-8 形式で保存されている必要がある)．この際，いわゆるリプライは含むが，リツイートは含まない設定にしているが，逆に一定数以上のリツイートがあったものだけを収集することも可能である．

なお，どのような検索ワードであっても，宣伝業者による同内容の莫大な量のツイートなど，分析に含むには適さないデータが多数含まれているため，不要なものは削除するなどさまざまな事後処理が必要である．

3. テキストマイニングの方法

https://github.com/ikegami-yukino/mecab/releases などを参照し，形態素分析ソフト MeCab をダウンロードして利用できるようにしたうえで Path を通す．これでテキストマイニングができるようになっているはずだが，MeCab の辞書は，ネットで使われるような新しい言葉が入っていないため，本研究では，新語を定期的に更新している mecab-ipadic-Neologd を Linux 経由でダウンロードしてデフォルトの辞書の代わりに用いた．

4. 結果と考察

2018 年 6 月 28 日 20 時 27 分にツイートの収集を行ったところ，6 月 18 日 22 時 06 分以降の「コーデ」を含むツイート 54383 個が収集された．これに，7 月 5 日 21 時 22 分に追加の収集を行い，さらに 42514 個収集し，接合した．目視で業者による広告を発見し，一括削除を行い，最終的に 88916 個を分析の対象にした．

図 1 は，この 17 日間の「コーデ」を含んだツイートの件数を 1 時間ごとに示したものである．おおむね同じパターであることが読み取れ，9 時から 12 時の間に小さなピーク，そして 20 時から 11 時くらいの間に大きなピークがあることが読み取れる．これは，午前中，人と会うなかで装いに関するツイートが増え，また夜，翌日以降の装いを考える中で増えているのかも知れない．一般的にツイート件数は早朝 4 時～5 時に最も少なく，右肩上がりで，21 時頃に多くなるという指摘(株式会社 AutoScale)と比較すれば，午前中に小さなピークが来るのは，「コーデ」の特徴といえるかも知れない．また，24 日 (2411)と 1 日 (0111)周辺だけ他の日と様相がことなり，11 時頃にピークが来ていることも，この日が日曜であることを考慮すると，この特徴の存在を支持するものといえよう．

表 1 はこの期間に「コーデ」とともにつぶやかれた名詞・形容詞のうち意味が取れる言葉の上位 25 語を示したものである．名詞は「今日」が最も多く，また近年の流行を反映して「双子」が 2 番目に多くなっている．また「大人」が 17 番目に入っている．形容詞は，「可愛い」，「良い」，「うれしい」などポジティブな言葉が並び，「かっこいい」，「大人っぽい」など少し詳しく説明する言葉がそれに続いている．

また，図2は，コーデとともにつぶやかれる名詞7つ（今日・双子私・夏・ファッション・スカート・大人）について，その出現頻度の日ごとの変動を表したものである．ここでも日曜日には平日とは異なるパターンが見られ，7月1日は「スカート」が増えて「大人」が減っている．これは「コーデ」という言葉を使う人々の価値観を反映したものといえるだろう．

5. 今後の課題

本研究は，Python と MeCab を使って，「コーデ」を含むツイートの分析を試みた．その結果，「コーデ」およびそれと一緒に用いられる言葉の時間的な変化をとらえることができた．しかし，本研究には様々な問題点がある．第1に，本研究が対象とした時間は短い．そのためその変化の多くは，言語使用の変化をとらえるのには不十分であり，その日の天候や出来事などに大きく依存している．今後は，より長期間のデータを集積したうえで，「コーデ」をめぐる表現がどのように変化するかを検討する必要があろう．第2に，本研究では，独自の辞書登録は行わず，Neologd の辞書そのままで形態素分析を行っている．そのため，誤分類と思われるものもときには見られる．また，自動化した処理で行われていることによる誤分類や，発話意図が他と大きく異なるツイートを含んでいることによる誤分類もある．これらは結果をゆがめる可能性のあるものではあるので，注意が必要である．

謝辞：
Python や MeCab による Twitter の分析に当たっては，mecab-ipadic-Neologd を含め，それぞれのプログラム，辞書の開発者はもちろん，ウェブ上の情報提供者にその多くを頼っている．ここで感謝の意を表する．

参考文献

藤田秀之 (2011). Twitter メッシュデータ収集・視覚化システム 地理情報システム学会第 20 回研究大会講演論文集, (https://www.gisa-japan.org/conferences/proceedings/2011/papers/E-6-3.pdf 2018 年 7 月 5 日確認)

五味伸之・辰巳暢・新田優喜 (2012). Twitter を利用した言語形態の変化についての研究 福井工業高等専門学校 研究紀要 人文・社会科学, 45, 1-6.

株式会社 AutoScale (2017). 【Twitter 運用ツール「Cheetah（チーター）」が 100 万ツイートを対象に調査】ツイートが拡散されやすい投稿時間帯は 5 時，11 時・15 時 (https://prtimes.jp/main/html/rd/p/000000004.000022240.html 2018 年 7 月 6 日確認)

北村智・河井大介・佐々木裕一 (2017). ソーシャルメディアにおける感情語使用と投稿動機，ネットワーク構造の関係 ―ツイッターでのポジティブ感情語・ネガティブ感情語に着目して 社会言語科学, 20, 16-28.

岡田祥平・西川由樹 (2016). 日本語研究資料としての Twitter ： コミュニケーション構造の観点から 新潟大学教育学部研究紀要 人文・社会科学編, 9, 93-111.

澤山郁夫・三宅幹子 (2018). 大学生の独り言的ツイートは独り言なのか―発話傾向との関連から パーソナリティ研究, 27, 31-41.

保田祥・岡本雅史・荒牧英治 (2012). <新しさ>のために循環する表現－女性向けファッション雑誌『InRed』を材料に 社会言語科学会第 29 回大会発表論文集, 134-137.

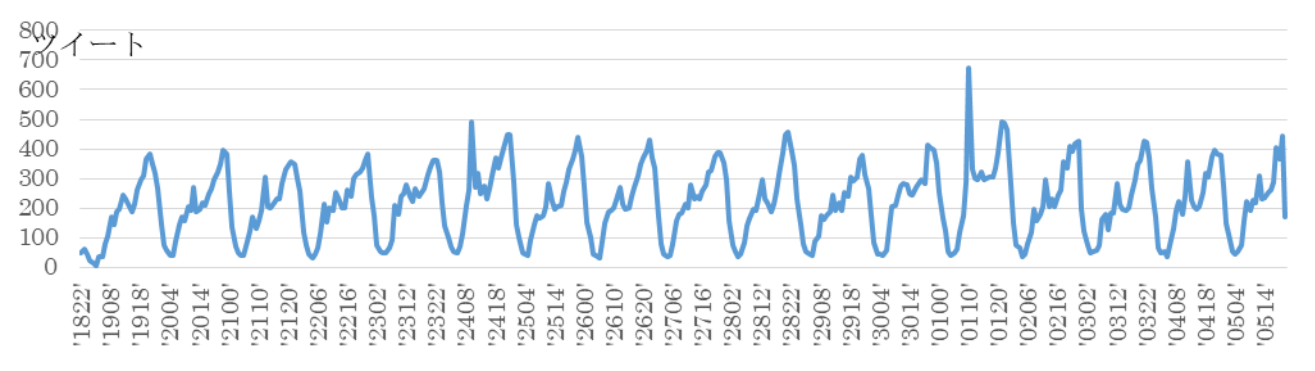

図１　「コーデ」を含んだツイートの時間変動（横軸は日と時刻：例えば 1822 は 18 日 22 時）

表１　「コーデ」とともにつぶやかれた名詞・形容詞のうち意味が取れる言葉の上位 25 語

名詞				形容詞			
今日	7473	こちら	2449	可愛い	7742	多い	957
双子	5636	スカート	2384	いい	4052	暑い	868
夏	4993	黒	2375	ない	4022	かっこいい	661
私	4260	大人	2361	かわいい	3243	難しい	653
好き	4189	シャツ	2154	良い	2558	高い	609
ファッション	3145	目	2153	嬉しい	2262	大人っぽい	580
服	3075	オシャレ	2130	欲しい	2198	早い	519
人	2706	おしゃれ	2108	楽しい	1745	無い	509
これ	2630	白	2075	すごい	1437	詳しい	504
色	2597	時	2056	よい	1396	やばい	497
交換	2518	T シャツ	1954	っぽい	1189	新しい	489
素敵	2514	画像	1929	ほしい	1008	おそい	467
アイテム	2461			やすい	1000		

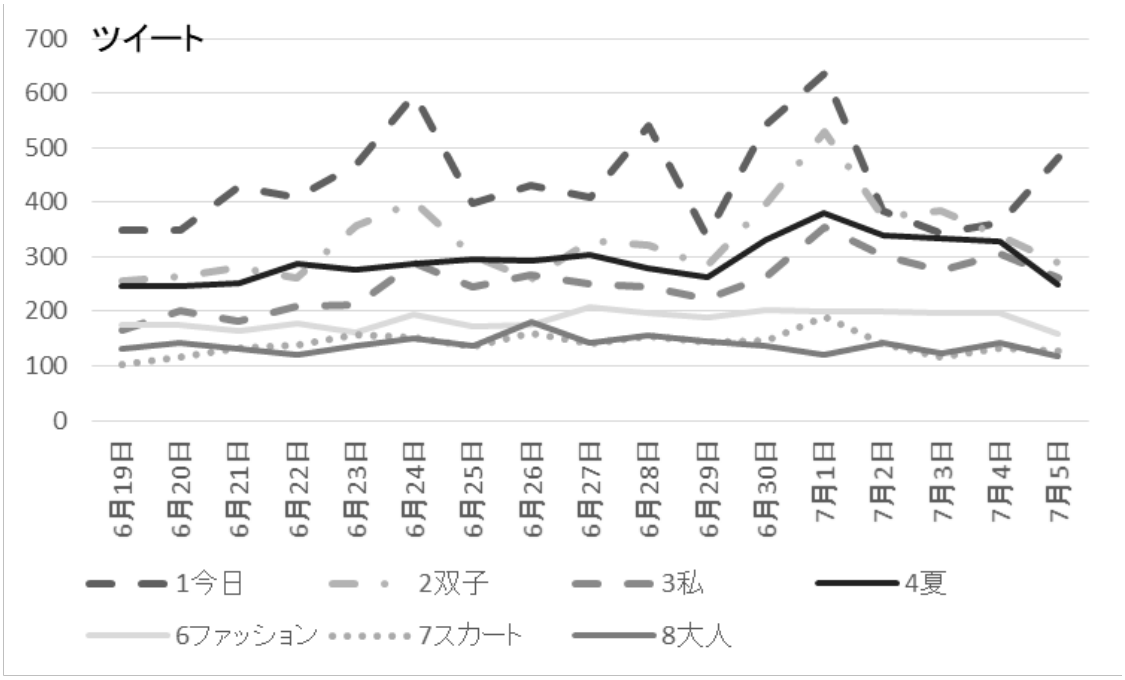

図２　「コーデ」とともにつぶやかれるキーワードの変動